IN THE UNITED STATES DISTRICT COURT
FOR THE NORTHERN DISTRICT OF CALIFORNIA
OAKLAND DIVISION

JOHN DOE I, et al., on behalf of themselves and all others similarly situated,

Plaintiffs,

v.

GOOGLE LLC,

Defendant.

Case No. 5:23-cv-02431-BLF

DECLARATION OF DR. ZUBAIR SHAFIQ IN SUPPORT OF PLAINTIFFS' MOTION FOR PRELIMINARY INJUNCTION

DECLARATION OF DR. ZUBAIR SHAFIQ

I, ZUBAIR SHAFIQ, hereby declare under penalty of perjury:

1. I am an associate professor of computer science at University of California, Davis, where I teach, conduct research and analyze issues involving Internet privacy, security, and performance, using network measurement and machine learning techniques.

2. I have been retained by Plaintiffs' counsel and submit this declaration in support of Plaintiffs' Motion for Preliminary Injunction.

3. I have personal knowledge of the facts set forth herein and, if called as a witness, could and would testify competently to them.

4. I reserve the right to modify, supplement or otherwise amend my statements, analyses, and conclusions in this declaration should new and additional information become available to me.

I. QUALIFICATIONS

1. I am an associate professor of computer science at University of California, Davis, where I run a lab and conduct research that focuses on Internet privacy, security, and performance, using network measurement and machine learning techniques. In particular, my research over the last several years has specifically aimed to uncover personal data collection, sharing, and usage in the online advertising ecosystem. In addition to my lab and research work, I regularly teach undergraduate and graduate courses on computer networks and computer security, including special topics courses covering emerging trends in online advertising and tracking.

2. I have co-authored nearly 100 peer-reviewed research papers, and have received several awards and distinctions for my research. Notably, I was the recipient of the 2018 Andreas Pfitzmann Award at the flagship Privacy Enhancing Technologies Symposium for my research on designing a system to reliably detect advertising and tracking information flows in mobile apps. I also received the Best Paper Award at the 2017 ACM Internet Measurement Conference for my research on exposing and investigating the abuse of a security vulnerability in Facebook Graph API's implementation of 3rd party apps. I also received the Best Paper Award at the 2012 IEEE

International Conference on Network Protocols for my research on reverse engineering proprietary network protocols through network traffic analysis.

3. My full qualifications are set forth in my curriculum vitae, which is attached as Exhibit A.

4. I have been retained by counsel for Plaintiffs as an expert in this matter and submit this declaration in support of Plaintiffs' Motion for Preliminary Injunction.

5. I reserve the right to amend, modify, or supplement my opinions as new or additional information becomes available to me

II. ASSIGNMENT AND SUMMARY OF CONCLUSIONS

6. I have been asked to conduct entropy analysis to determine whether the information obtained by Google source code – specifically Google Analytics, Google Ads, and Google Display Ads – may be considered personally identifiable information (PII) about patients on health care related websites and apps. Based on my analysis of the information collected, my conclusion is yes. As I explain below, the information collected by Google far exceeds a critical amount of user-identifying entropy. The information collected by Google allows it to identify patients uniquely and persistently on health care related websites and apps.

7. I have been asked to analyze whether a data company, like Google, can readily identify health care related websites and apps. My conclusion is yes. Data companies, like Google, should have the systems and tools in place to identify website and app content. In the case of Google, I am aware that Google does indeed have specific tools that can do this work. Based on my experience and expertise, as well as my analysis of various tools available to data companies, it is my opinion that Google can identify health care related websites and apps using its own Verticals taxonomy and content classification API, as well as off-the-shelf content classification services.

III. OVERVIEW OF ENTROPY

8. The scientific community has widely adopted entropy as a metric to quantify privacy risk of identifiability of an individual.[1] Entropy is measured in terms of bits. If the number of entropy bits for a piece of data – either alone or combined with other information – exceed a certain threshold, then the concept of entropy concludes that the data can be used to uniquely (re)identify users.

9. Google itself uses entropy to quantify privacy risk. For example, Google's Privacy Sandbox project uses entropy to determine "Privacy Budget."[2] Outside of Google, the privacy non-profit public interest group Electronic Frontier Foundation[3] and the Mozilla[4] browser also used entropy as a metric to assess identifiability of information.

10. Google Chrome uses entropy to label APIs "that exposes data that folks on the internet find useful for fingerprinting". "Attributes and methods marked as [HighEntropy] are

[1] Note that dozens of privacy metrics and analysis techniques have been proposed in the scientific community over the past few decades. The following paper by Wagner and Eckhoff – that surveys more than 80 privacy metrics – explains that most of these metrics either directly build on entropy or are indirectly related to entropy. For this reason, and – as explained below – because Google also extensively uses entropy for privacy assessment, I use entropy for privacy analysis. However, the same conclusion can be reached if other privacy metrics are suitably used to analyze the information collected by the Google Source Code – specifically Google Analytics, Google Ads, and Google Displays Ads.

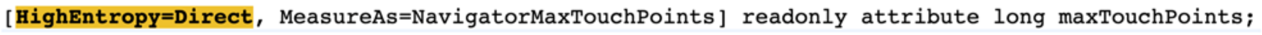
Wagner, I. and Eckhoff, D., 2018. Technical privacy metrics: a systematic survey. ACM Computing Surveys, 51(3).

[2] Privacy Budget: Limit the amount of individual user data exposed to sites to prevent covert tracking. https://developer.chrome.com/en/docs/privacy-sandbox/privacy-budget/

[3] A Primer on Information Theory and Privacy https://www.eff.org/deeplinks/2010/01/primer-information-theory-and-privacy

[4] Technical Comments on Privacy Budget https://mozilla.github.io/ppa-docs/privacy-budget.pdf

known to be practically useful for identifying particular clients on the web today.”[5] See, for example[6] (emphasis added):



```
[HighEntropy=Direct, MeasureAs=NavigatorMaxTouchPoints] readonly attribute long maxTouchPoints;
```

11. The minimum amount of entropy required to uniquely identify a person in a population of size N is $\log_2(N)$. Given that Earth’s population is approximately 8 billion, the number of required bits is $\log_2$ (8 billion) = 32.897 ≈ 33 bits. Given that the number of Internet users on Earth is ≈ 4 billion, the number of required entropy bits to uniquely identify a user or device on the Internet is log2(4 billion) = 31.897 ≈ 32 bits.[7]

12. Google itself uses 32 bits of entropy as the identifiability threshold. As shown in Figure 1, Google uses the 32 bits as the identifiability threshold[8] in calculating its “Privacy Budget”. As another example, as shown in Figure 2, the FAQ page of Google’s Privacy Budget project justifies the use of the 32-bit entropy threshold for identifiability.[9]

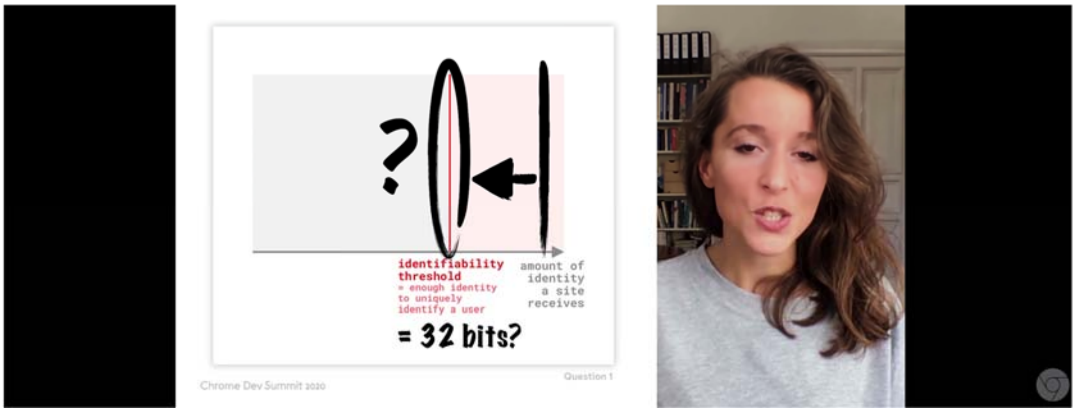

?
identifiability threshold
= enough identity to uniquely identify a user
amount of identity a site receives
= 32 bits?
Chrome Dev Summit 2020
Question 1

Figure 1: Google employee explains that the 32-bit identifiability threshold is enough to uniquely identify a user.

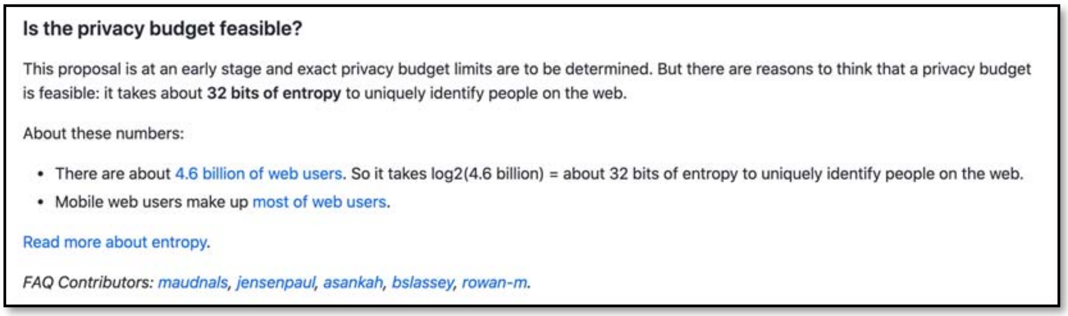
[5] https://chromium.googlesource.com/chromium/src/+/main/third_party/blink/renderer/bindings/IDLExtendedAttributes.md#HighEntropy_m_a_c

[6] https://source.chromium.org/search?q=HighEntropy%3DDirect&ss=chromium%2Fchromium%2Fsrc

[7] https://www.eff.org/deeplinks/2010/01/primer-information-theory-and-privacy

[8] Introducing the Privacy Budget https://www.youtube.com/watch?v=0STgfjSA6T8&t=423s

[9] https://github.com/mikewest/privacy-budget/blob/4e5f78adde92bd622dafeceae78682fc0823c0eb/faq.md



Is the privacy budget feasible?

This proposal is at an early stage and exact privacy budget limits are to be determined. But there are reasons to think that a privacy budget is feasible: it takes about **32 bits of entropy** to uniquely identify people on the web.

About these numbers:

- There are about 4.6 billion of web users. So it takes log2(4.6 billion) = about 32 bits of entropy to uniquely identify people on the web.
- Mobile web users make up most of web users.

Read more about entropy.

FAQ Contributors: maudnals, jensenpaul, asankah, bslassey, rowan-m.

Figure 2: Google explains entropy and the 32-bit identifiability threshold.

13. Thus, under the concept of entropy, individual pieces of data in a given transmission are evaluated and assigned a value of bits. The bits are then totaled and if they exceed 32-bits then that information combination is deemed to be identifiable within the scientific community.[10]

IV. ANALYSIS

14. To apply the entropy concept, I analyzed network transmissions from an individual's web browser when they are communicating with a health care provider, for example, MedStar Health. I then focus on the network transmissions that, in the course of the individual's communication with MedStar, are also being made to Google via Google Analytics, Google Ads and Google Displays Ads. Specifically, I look at the network transmissions that occur in the following example: when loading the "Abdominal Aneurysm Treatment | MedStar Health" webpage included in the Amended Complaint: https://www.medstarhealth.org/services/abdominal-aneurysm-treatment.

15. I found that Google collected at least the following four types of identifying information from the patient's browser:[11]

[10] A caveat to be aware of when calculating the "joint" entropy is that we should not simply sum up entropy of different pieces of data if they are dependent with each other. If different pieces of data are dependent, then the joint entropy could be lower than the simple sum of entropy of different pieces of data. Thomas M. Cover; Joy A. Thomas. Elements of Information Theory. Wiley (2006)

[11] I focus on these four types of information because, as discussed further below, they far exceed the entropy threshold for identifiability. But, based on my experience, expertise, and review of network transmissions, there is additional information that is sent to Google, which may also contribute to the entropy threshold. For the sake of brevity, I do not analyze each and every piece of data transmitted but can do so if the Court so requires.

a. IP address

b. User agent

c. Cookies, such as:[12]

i. _ga cookie used to distinguish patients with 2 years' expiration and exfiltrated via cid URL parameter;

ii. _gid cookie used to distinguish patients with 24 hours' expiration and exfiltrated via _gid URL parameter;

iii. _gcl_au cookie used for conversion tracking with 3 months' expiration and exfiltration via auiddc URL parameter;

iv. Third-party AEC cookie on google.com with session based expiration;

v. Third-party IDE cookie on doubleclick.net containing unique device identifier with 13 months' expiration;

vi. Third-party DSID cookie on doubleclick.net containing Google account identifier with 2 weeks' expiration; and

vii. Third-party NID cookie on google.com containing unique device identifier with 6 months' expiration.

d. Device attributes, such as:

i. Screen resolution exfiltrated via sr URL parameter;

ii. Viewport size exfiltrated via vp URL parameter;

iii. Document encoding exfiltrated via de URL parameter;

iv. Screen color depth exfiltrated via sd URL parameter;

v. Language exfiltrated via ul URL parameter;

vi. Whether Java is enabled exfiltrated via je URL parameter;

vii. Device architecture exfiltrated via uaa URL parameter;

viii. Device bitness exfiltrated via uab URL parameter;

[12] I note that while the _ga, _gid, and _gcl_au cookies belong to Google (a third party to the communication between the individual and their health care provider), these three Google cookies nonetheless appear as "first-party" cookies. This makes it much more difficult for individuals to prevent the transmission of information to third-party Google.

ix. Whether the device is mobile exfiltrated via uamb URL parameter;

x. Device model exfiltrated via uam URL parameter;

xi. Device platform exfiltrated via uap URL parameter;

xii. Device platform version number exfiltrated via uapv URL parameter;

xiii. Whether Windows OS on Windows 64 bit is supported exfiltrated via WOW64 parameter;

xiv. User agent full version list exfiltrated via uafvl URL parameter;

xv. Accepted encoding via Accept-Encoding header;

xvi. Accepted language via Accept-Language header;

xvii. Cache control via Cache-Control header; and

xviii. Which content types client is able to understand via Accept header.

16. Below I conduct entropy analysis of the aforementioned four types of identifying information collected by Google.

17. First, the IP address by itself is a sufficiently unique and persistent identifier to be classified as PII.

a. There are two prevalent IP protocols: IPv4 and IPv6. The length of IPv4 address is 32 bits (i.e., ≈4 billion possible IPv4 addresses). Note that IPv4 addresses are sometimes reused or shared across users (e.g., using a mechanism called Network Address Translation [NAT]). The new IPv6 protocol is now used by approximately one-half of Internet users in the United States.[13] The length of IPv6 address is 128 bits (i.e., ≈340 trillion-trillion-trillion IP addresses possible IPv6 addresses). Thus, IP address, especially the newer IPv6 variant, is able to uniquely identify patients.

b. While IP addresses may not always be static (i.e., they can change), peer-reviewed research[14] shows that the IP address by itself remains a serious threat to tracking despite the use of non-static IP addresses. Specifically, researchers showed that

[13] https://www.google.com/intl/en/ipv6/statistics.html

[14] Don’t Count Me Out: On the Relevance of IP Address in the Tracking Ecosystem https://dl.acm.org/doi/pdf/10.1145/3366423.3380161

"87% of participants retain at least one IP address for more than a month". For the study participants in the United States, the average IP address retention period was 18.93 days. Thus, IP address is a persistent identifier.

18. Second, the user agent also contains a significant amount of entropy. There are approximately 10 bits of entropy in user agent. More specifically, 10.000 according to this EFF study[15] and 9.779 bits according to this AmIUnique study.[16] As discussed below, user agent when combined with other information collected by Google easily exceeds the 32-bit identifiability threshold.

19. Third, since arbitrary information can be stored in cookies, there is really no limit to how many bits of entropy (identifying information) can be stored in cookies. Google typically stores Universally Unique Identifier (UUID) in the aforementioned cookies.[17] Assuming each character encodes up to 8 bits of information, the total entropy for the NID cookie alone would be 211 characters x 8 bits = 1,688 bits. Similarly, each of the other identifier cookies listed above far exceeds the 32-bit identifiability threshold.

20. Fourth, the various device attributes collected by Google also contain sufficient entropy that can be combined[18] with IP address and user agent to exceed the 32-bit identifiability threshold. For example, IP address (containing up to 32 bits for IPv4 and up to 128 bits for IPv6) and user agent (containing approximately 10 bits) can be combined with various device attributes: screen/viewport size and color depth (7.72 bits), screen resolution (4.89 bits), Accept-Language

[15] Peter Eckersley. How unique is your web browser? International Symposium on Privacy Enhancing Technologies Symposium, pages 1–18. Springer, 2010.

[16] Pierre Laperdrix, Walter Rudametkin, and Benoit Baudry. Beauty and the beast: Diverting modern web browsers to build unique browser fingerprints. IEEE Symposium on Security and Privacy, pages 878–894. 2016.

[17] https://developers.google.com/analytics/devguides/collection/protocol/v1/parameters#cid

[18] While the overall entropy of a combination of fields cannot be computed by simply adding them up if they are not independent as I discussed above, there is clear sufficient entropy together in these fields such that the joint entropy easily surpasses the 32-bit identifiability threshold.

(5.918 bits), Accept-Encoding (1.53 bits), Device platform (2.31 bits), Accept (2.21 bits), and whether device is mobile (0.68 bits) to exceed the 32 bit identifiability threshold.[19,20]

21. In summary, the various pieces of information collected by Google Analytics, Google Ads, and Google Displays Ads far exceeds the 32-bits entropy identifiability threshold. My conservative analysis shows that the information collected by Google allows it to identify individuals uniquely and persistently on health care related websites and apps.

V. HEALTH CARE CONTENT CLASSIFICATION

22. Google provides advertisers a well-known "Verticals" taxonomy[21] to target content of various categories. These include at least 89 specific "/Health" related verticals such as "/Health/Health Conditions/Cancer" and "/Health/Health Conditions/AIDS & HIV".

23. It is my opinion that Google possesses the technology to use its taxonomies to classify content on health care related websites and apps. Google's initial claim to fame was its search engine that is based on "processing and analyzing the textual content and key content tags and attributes".[22]

24. Based on my experience and expertise, it is my opinion that Google can likely easily identify health care related websites and apps using this content classification technology. In fact, Google publicly provides a "content categories" classification API[23] that it can use to identify health care related websites and apps, including those that are classified into verticals such as "/Health/Health Conditions/Cancer" and "/Health/Health Conditions/AIDS & HIV".

[19] https://coveryourtracks.eff.org/

[20] Pierre Laperdrix, Walter Rudametkin, and Benoit Baudry. 2016. Beauty and the beast: Diverting modern web browsers to build unique browser fingerprints. 37th IEEE Symposium on Security and Privacy.

[21] https://developers.google.com/adwords/api/docs/appendix/verticals

[22] https://developers.google.com/search/docs/fundamentals/how-search-works

[23] Cloud Natural Language API https://cloud.google.com/natural-language/docs/categories

25. In addition, I am aware of several off-the-shelf content classification services (e.g., WebShrinker[24], RapidAPI[25], Klazify[26]) that Google and anyone else can use to identify health care related websites and apps. These services are available for free (but rate and quota limited) or can be obtained at a reasonable cost (e.g., $100 per month for 30 million content classification decisions[27]).

* * *

I declare under penalty of perjury under the laws of the United States of America that the foregoing is true and correct.

Executed this 12th day of June 2023 at Davis, California.

/s/ Zubair

Dr. Zubair Shafiq

[24] Website Categories https://docs.webshrinker.com/v3/iab-website-categories.html#tier-1-and-tier-2-categories

[25] https://rapidapi.com/ibmbpmtips/api/iab-taxonomy-text-classification/details

[26] https://www.klazify.com/

[27] https://rapidapi.com/ibmbpmtips/api/iab-taxonomy-text-classification/pricing